APPENDIX

A. PROOFS

A.1 Proof of Statement 1

PROOF. Each coordinate $N \cdot s'_{l'}$ of the vector in (4) is, by definition of partial supports, just the number of transactions in the randomized sequence T' that have intersections with A of size l'. Each randomized transaction t' contributes to one and only one coordinate $N \cdot s'_{l'}$, namely to the one with $l' = \#(t' \cap A)$. Since we are dealing with a per-transaction randomization, different randomized transactions contribute independently to one of the coordinates. Moreover, by item-invariance assumption, the probability that a given randomized transaction contributes to the coordinate number l' depends only on the size of the original transaction t (which equals m) and the size l of intersection $t \cap A$. This probability equals $p\,[l \to l']$.

So, for all transactions in T that have intersections with A of the same size l (and there are $N \cdot s_l$ such transactions) the probabilities of contributing to various coordinates $N \cdot s'_{l'}$ are the same. We can split all N transactions into $k+1$ groups according to their intersection size with A. Each group contributes to the vector in (4) as a multinomial distribution with probabilities

$$(p\,[l \to 0]\,, p\,[l \to 1]\,, \ldots\,, p\,[l \to k]),$$

independently from the other groups. Therefore the vector in (4) is a sum of $k+1$ independent multinomials. Now it is easy to compute both expectation and covariance.

For a multinomial distribution $(X_0, X_1, \ldots, X_k)$ with probabilities $(p_0, p_1, \ldots, p_k)$, where $X_0 + X_1 + \ldots + X_k = n$, we have $\mathbf{E}\,X_i = n \cdot p_i$ and

$$\mathrm{Cov}\,(X_i, X_j) = \mathbf{E}\,(X_i - p_i)(X_j - p_j) = n \cdot (p_i \delta_{i=j} - p_i p_j).$$

In our case, X_i = l's part of $N \cdot s'_i$, $n = N \cdot s_l$, and $p_i = p\,[l \to i]$. For a sum of independent multinomial distributions, their expectations and covariances add together:

$$\mathbf{E}\,(N \cdot s'_{l'}) = \sum_{l=0}^{k} N \cdot s_l \cdot p\,[l \to l']\,,$$

$$\mathrm{Cov}\,(N \cdot s'_i, N \cdot s'_j) =$$
$$= \sum_{l=0}^{k} N \cdot s_l \cdot (p\,[l \to i] \cdot \delta_{i=j} - p\,[l \to i] \cdot p\,[l \to j])$$

Thus, after dividing by an appropriate power of N, the formulae in the statement are proven. □

A.2 Proof of Statement 2

PROOF. We are given a transaction $t \in T$ and an itemset $A \subseteq \mathcal{I}$, such that $|t| = m$, $|A| = k$, and $\#(t \cap A) = l$. In the beginning of randomization, a number j is selected with distribution $\{p_m[j]\}$, and this is what the first summation takes care of. Now assume that we retain exactly j items of t, and discard $m-j$ items.

Suppose there are q items from $t \cap A$ among the retained items. How likely is this? Well, there are $\binom{m}{j}$ possible ways to choose j items from transaction t; and there are $\binom{l}{q}\binom{m-l}{j-q}$ possible ways to choose q items from $t \cap A$ and $j-q$ items from $t \setminus A$. Since all choices are equiprobable, we get $\binom{l}{q}\binom{m-l}{j-q} / \binom{m}{j}$ as the probability that exactly q A-items are retained.

To make t' contain exactly l' items from A, we have to get additional $l'-q$ items from $A \setminus t$. We know that $\#(A \setminus t) = k - l$, and that any such item has probability ρ to get into t'. The last terms in (8) immediately follow. Summation bounds restrict q to its actually possible (= nonzero probability) values. □

A.3 Proof of Statement 3

PROOF. Let us denote

$$\vec{p}_l := (p\,[l \to 0]\,, p\,[l \to 1]\,, \ldots\,, p\,[l \to k])^T,$$
$$\vec{q}_l := (q\,[l \leftarrow 0]\,, q\,[l \leftarrow 1]\,, \ldots\,, q\,[l \leftarrow k])^T.$$

Since $PQ = QP = I$ (where I is the identity matrix), we have

$$\sum_{l=0}^{k} p\,[l \to i]\; q\,[l \leftarrow j] = \sum_{l'=0}^{k} p\,[i \to l']\; q\,[j \leftarrow l'] = \delta_{i=j}.$$

Notice also, from (7), that matrix $D[l]$ can be written as

$$D[l] = \mathrm{diag}(\vec{p}_l) - \vec{p}_l\, \vec{p}_l^{\;T},$$

where $\mathrm{diag}(\vec{p}_l)$ denotes the diagonal matrix with $\vec{p}_l$-coord-

inates as its diagonal elements. Now it is easy to see that

$$\tilde{s} = \vec{q}_k{}^T \vec{s}\,' = \sum_{l'=0}^{k} q\,[k \leftarrow l'] \cdot s'_{l'};$$

$$\begin{aligned}
\mathrm{Var}\ \tilde{s} &= \frac{1}{N} \sum_{l=0}^{k} s_l\, \vec{q}_k{}^T D[l]\, \vec{q}_k = \\
&= \frac{1}{N} \sum_{l=0}^{k} s_l\, \vec{q}_k{}^T \left(\mathrm{diag}(\vec{p}_l) - \vec{p}_l \vec{p}_l{}^T\right) \vec{q}_k = \\
&= \frac{1}{N} \sum_{l=0}^{k} s_l \left(\vec{q}_k{}^T \mathrm{diag}(\vec{p}_l)\, \vec{q}_k - (\vec{p}_l{}^T \vec{q}_k)^2\right) = \\
&= \frac{1}{N} \sum_{l=0}^{k} s_l \Big(\sum_{l'=0}^{k} p\,[l \to l']\ q\,[k \leftarrow l']^2 - \delta_{l=k} \Big);
\end{aligned}$$

$$\begin{aligned}
&(\mathrm{Var}\ \tilde{s})_{est} \ = \\
&= \frac{1}{N} \sum_{l=0}^{k} (\vec{q}_l{}^T \vec{s}\,') \Big(\sum_{l'=0}^{k} p\,[l \to l']\ q\,[k \leftarrow l']^2 - \delta_{l=k} \Big) = \\
&\quad = \frac{1}{N} \sum_{j=0}^{k} s'_j \Big(\sum_{l,l'=0}^{k} q\,[l \leftarrow j]\ p\,[l \to l']\ q\,[k \leftarrow l']^2 - \\
&- \sum_{l=0}^{k} \delta_{l=k}\, q\,[l \leftarrow j] \Big) \ = \ \frac{1}{N} \sum_{j=0}^{k} s'_j \Big(\sum_{l'=0}^{k} \delta_{l'=j}\, q\,[k \leftarrow l']^2 - \\
&\qquad - q\,[k \leftarrow j] \Big) \ = \ \frac{1}{N} \sum_{j=0}^{k} s'_j \big(q\,[k \leftarrow j]^2 - q\,[k \leftarrow j] \big).
\end{aligned}$$

□

A.4 Proof of Statement 4

PROOF. We prove the left formula in (13) first, and then show that the right one follows from the left one. Consider $N \cdot \Sigma_l$; it equals

$$N \cdot \Sigma_l = N \cdot \sum_{C \subseteq A,\ |C| = l} \mathrm{supp}^T(C) = \sum_{C \subseteq A,\ |C| = l} \#\{t_i \in T \mid C \subseteq t_i\} = \\ = \sum_{i=1}^{N} \#\{C \subseteq A \mid |C| = l, C \subseteq t_i\}.$$

In other words, each transaction t_i should be counted as many times as many different l-sized subsets $C \subseteq A$ it contains. From simple combinatorics we know that if $j = \#(A \cap t_i)$ and $j \geqslant l$, then t_i contains $\binom{j}{l}$ different l-sized subsets of A. Therefore,

$$N \cdot \Sigma_l = \sum_{i=1}^{N} \binom{\#(A \cap t_i)}{l} = \\ = \sum_{j=l}^{k} \binom{j}{l} \cdot \#\{t_i \in T \mid \#(A \cap t_i) = j\} = \sum_{j=l}^{k} \binom{j}{l} N \cdot s_j,$$

and the left formula is proven. Now we can check the right formula just by replacing the Σ_j's according to the left formula. We have:

$$\begin{aligned}
&\sum_{j=l}^{k} (-1)^{j-l} \binom{j}{l} \Sigma_j = \sum_{j=l}^{k} (-1)^{j-l} \binom{j}{l} \sum_{q=j}^{k} \binom{q}{j} s_q = \\
&= \sum_{l \leqslant j \leqslant q \leqslant k} (-1)^{j-l} \binom{j}{l} \binom{q}{j} s_q = \sum_{q=l}^{k} s_q \sum_{j=l}^{q} (-1)^{j-l} \binom{j}{l} \binom{q}{j} \\
&\quad = \sum_{q=l}^{k} s_q \sum_{j'=0}^{q-l} (-1)^{j'} \frac{(j'+l)!}{l!\, j'!} \frac{q!}{(j'+l)!\,(q-j'-l)!} = \\
&\quad = \sum_{q=l}^{k} s_q \cdot \frac{q!}{l!\,(q-l)!} \sum_{j'=0}^{q-l} (-1)^{j'} \frac{(q-l)!}{j'!\,(q-l-j')!} = \\
&\qquad\qquad = \sum_{q=l}^{k} s_q \binom{q}{l} \sum_{j'=0}^{q-l} (-1)^{j'} \binom{q-l}{j'} = s_l,
\end{aligned}$$

since the sum $\sum_{j'=0}^{q-l} (-1)^{j'} \binom{q-l}{j'}$ is zero whenever $q - l > 0$.

To prove that matrix P becomes lower triangular after the transformation from $\vec{s}$ and $\vec{s}\,'$ to $\vec{\Sigma}$ and $\vec{\Sigma}\,'$, let us find how $\mathbf{E}\ \vec{\Sigma}\,'$ depends on $\vec{\Sigma}$ using the definition (12).

$$\begin{aligned}
&\mathbf{E}\ \Sigma'_{l'} = \sum_{C \subseteq A,\ |C| = l'} \mathbf{E}\ \mathrm{supp}^{T'}(C) = \\
&\quad = \sum_{C \subseteq A,\ |C| = l'} \sum_{l=0}^{l'} p_{l'}^{m}\,[l \to l'] \cdot \mathrm{supp}_l^T(C) = \\
&= \sum_{C \subseteq A,\ |C| = l'} \sum_{l=0}^{l'} p_{l'}^{m}\,[l \to l'] \sum_{j=l}^{l'} (-1)^{j-l} \binom{j}{l} \Sigma_j(C,T) = \\
&= \sum_{j=0}^{l'} \underbrace{\sum_{l=0}^{j} (-1)^{j-l} \binom{j}{l} p_{l'}^{m}\,[l \to l']}_{c_{l'j}} \sum_{C \subseteq A,\ |C| = l'} \Sigma_j(C,T) = \\
&\quad = \sum_{j=0}^{l'} c_{l'j} \sum_{C \subseteq A,\ |C| = l'} \sum_{B \subseteq C,\ |B| = j} \mathrm{supp}^T(B) = \\
&= \sum_{j=0}^{l'} c_{l'j} \sum_{B \subseteq A,\ |B| = j} \#\{C \mid B \subseteq C \subseteq A, |C| = l'\} \cdot \mathrm{supp}^T(B) = \\
&= \sum_{j=0}^{l'} c_{l'j} \sum_{B \subseteq A,\ |B| = j} \binom{k-j}{l'-j} \mathrm{supp}^T(B) = \sum_{j=0}^{l'} c_{l'j} \binom{k-j}{l'-j} \cdot \Sigma_j.
\end{aligned}$$

Now it is clear that only the lower triangle of the matrix can have non-zeros. □